Evolutionary Inference for Function-valued Traits: Gaussian Process Regression on Phylogenies

Nick S. Jones[1] and John Moriarty[2]

[1]*Department of Mathematics, Imperial College London, SW7 2AZ*

[2]*School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, UK*

This paper combines concepts from phylogenetics with Gaussian process regression, in order to allow evolutionary inference for function-valued traits that are correlated through phylogeny. By function-valued traits, we mean data objects which are indexed by one or more 'spatial' co-ordinates, such as organism age or buffer pH level. Examples of function-valued traits include organism mass with age or fitness vs pH curves. We provide a nonparametric Bayesian model for such data by using Gaussian processes. The model may be used to infer ancestral function-valued traits, compare rates of evolution across a phylogeny, or to identify the most likely phylogenies consistent with observed data. It contrasts with methods which reduce data to summary statistics or a multivariate vector (without spatial co-ordinates) and allows us to make inferential statements about ancestral traits themselves. We illustrate the use of the method on real data we generated by experiment.

I. INTRODUCTION

In this paper we consider the problem of evolutionary inference for function-valued traits. By *evolutionary inference* we mean statistical inference for data objects related by a phylogenetic tree (more specifically, a chronogram), and we use the term *function-valued trait* to mean a data object with a spatial co-ordinate [1]. Since the phylogenetic tree will play the role of time, we refer to the co-ordinate of the trait as *space*. Example traits are curves for ambient temperature versus growth rate for caterpillars, heart rhythm time series [2], or spectrograms of audio data. Figure 1A gives an example of function-valued traits, related by a phylogenetic tree, which we generated by experiment. As an example application, suppose that we have spectrograms of the calls of several bat species. Using the methods described below, we may estimate the phylogeny of the call (which may differ from inferred genetic phylogenies); alternatively, given a known phylogeny, we may dare to make predictions about the sound of an ancestral bat call, or test for equal rates of evolution along the phylogeny.

In order to perform evolutionary inference on function-valued traits, one might attempt to summarize each trait by a symbolic sequence (e.g. the presence or absence of certain characters) [3] or by multivariate vectors of continuous characters or summary statistics [4, 5], and then employ existing methods for phylogenetic inference. However there are many possible ways to choose such a summary, which may not be mutually compatible at the level of statistical models; also, as yet unobserved traits may not possess a particular landmark used to generate the summary. In contrast, function-valued traits are functions of a spatial co-ordinate and this implies an ordering on their values which may be used to capture important patterns of variation in the data [6]. Moreover, since the map from trait to summary is many-to-one, the use of summary statistics only allows indirect inference for ancestral traits: the inverse problem of inferring the ancestral trait from the ancestral summary would still remain. In the following we shall adopt a nonparametric functional approach, which does not in principle require either summarizing or constraining function-valued data, and allows inference to be made about functions.

Because of their generality, flexibility and mathematical simplicity, there has been much recent interest in the use of Gaussian process priors for Bayesian nonparametric regression. The monograph [7] gives a comprehensive treatment of their use in univariate regression, and both spatial and spatial-temporal Gaussian process models have been considered in the area of geostatistics

[8, 9]. In this paper we extend spatial-temporal Gaussian process modelling to take account of a tree topology, in order to model correlation due to shared evolutionary history. The method provides an explicit posterior distribution for ancestral traits, given observations from traits at any set of positions along the phylogeny. It can also be used to make inferences about the phylogeny or the evolutionary process operating along it.

Our work is related to that of Felsenstein [4], which gives a method for comparative studies of real-valued traits corrected for phylogeny. It has become widely used, together with its extensions [5, 10], to infer correlations in the evolution of quantitative characters. Our model may be regarded as a Bayesian view of this method, extended in two directions. Firstly, each position on the phylogeny carries a function-valued (rather than real- or vector-valued) trait. Secondly, in [4] individual traits may be interpreted as observations of a Brownian motion indexed by a phylogeny: this is generalised to a space-time Gaussian process indexed by a phylogeny. By extending Felsenstein's model to function-valued traits, we hope to partially address his observation that "the difficulty [in the multivariate case] is that quantitative characters will evolve at different rates, and in a correlated fashion" [4].

II. PHYLOGENETIC GAUSSIAN PROCESSES

A. Inference with Gaussian Processes

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution [7]. As a palette of statistical models, the set of Gaussian processes is large and flexible. Two examples used in the comparative method are the familiar Brownian Motion (Weiner process) and the Ornstein-Uhlenbeck process [11]. They have found application in this context as suitable models for continuous characters diffusing through evolutionary time. Despite this dynamical description, however, samples from any Gaussian process also have a very simple statistical description, as follows. Suppose that a Gaussian process f is observed at a vector of co-ordinates L. Then the resulting vector of sample values $f(L)$ can be viewed as just one sample from a multivariate Gaussian distribution of dimension equal to $|L|$, the number of points of measurement: $f(L) \sim \mathcal{N}(0, \sigma(L, L))$. Here, $\sigma(L, L)$ is the matrix of the covariances between all pairs (ℓ_i, ℓ_j) of observation co-ordinates in L (where $\ell_i \in L$), and $\mathcal{N}$ represents the Gaussian distribution, its two arguments being the mean and covariance matrix. Throughout this paper and the supplementary material, and solely for ease of exposition, we make the common assumption that the distribution mean is zero. The only choice we have when specifying a Gaussian process is thus how the sample values covary, which is encoded by the covariance function σ. Since the only restriction is that σ must be a positive semidefinite and symmetric function, Gaussian processes provide flexible models for a wide class of random processes. In practice, one typically assumes a particular form for σ, which may depend on hyperparameters θ. These hyperparameters often have interpretations as characteristic length scales, strengths of interaction between coordinate directions, and/or error standard deviations. The log-likelihood of the sample $f(L)$ is then

$$\log p(f(L)|\theta) = f(L)^T \sigma(L, L, \theta) f(L) - \log(det(\sigma(L, L, \theta))) - \frac{|L|}{2} \log 2\pi, \tag{1}$$

where we have made the dependence of σ on θ explicit.

Bayesian prediction is analytically tractable when assuming a Gaussian process prior distribution for a random function f. We might be interested in making inferences about the unobserved values of our random function f at a vector M of co-ordinates, given samples at the co-ordinates L. The posterior distribution of the vector $f(M)$ given $f(L)$ is also Gaussian and of the form:

$$f(M)|f(L) \sim \mathcal{N}(A, B) \tag{2}$$

where

$$A = \sigma(M,L)\sigma(L,L)^{-1}f(L), \quad (3)$$
$$B = \sigma(M) - \sigma(M,L)\sigma(L,L)^{-1}\sigma(M,L)^T \quad (4)$$

and $\sigma(M,L)$ denotes the $|M| \times |L|$ matrix of the covariance function σ evaluated at all pairs $m_i \in M, l_j \in L$. From Eq. (3), the posterior mean vector A consists of linear combinations of the observations. Note that the posterior covariance matrix B, given by (4), is independent of the observations. Gaussian process regression is nonparametric in the sense that no assumption is made about the structure of the model: the more data gathered, the longer the vector $f(L)$, and the more intricate the posterior model for $f(M)$. This is the sense in which the data is not being summarized in order to perform inference.

Though the Wiener process is one-dimensional, Gaussian processes can have arbitrary dimension. In the context of our model, this is important because we can consider each point of observation ℓ as corresponding to a point (x,t) in both space and evolutionary time. Here we define space as the co-ordinate of a function-valued-trait, as above. As an example, if the spatial coordinate is one-dimensional then $f(L)$ would represent the values of a random two-dimensional function at various space-time co-ordinates. A slice through this function at a fixed value of t can be viewed as a one-dimensional function-valued trait at the fixed time t in its evolution (with straightforward generalizations to higher-dimensional traits). In the next section we will extend this view to processes on phylogenies.

B. Phylogenetic covariance function

We now provide some formal details for the construction of a Gaussian process model for function-valued traits indexed by a phylogeny, $\mathbf{T}$. Readers preferring examples early on might skim the following and move to the next subsection and then the illustrative applications.

We suppose each observation ℓ corresponds to a point $(x, \mathbf{t})$ in the *space-phylogeny* $S \times \mathbf{T}$: that is, $x \in S$ is the value under consideration of the spatial (trait) coordinate, and $\mathbf{t} \in \mathbf{T}$ is the point under consideration on the phylogeny ($\mathbf{t}$ is not just a time co-ordinate but also indicates a branch of the phylogeny). Our aim in the following is to build a model for the evolution of a function-valued trait along this branched phylogeny. We will do this by constructing a covariance function $\Sigma_{\mathbf{T}}(\ell_i, \ell_j)$ when each l_i is a point in $S \times \mathbf{T}$. There are two assumptions from which our covariance function will follow (it should be noted that these assumptions are consistent with [4]):

Assumption II.1. *Conditional on their common ancestors in the phylogenetic tree* $\mathbf{T}$*, any two traits are statistically independent.*

Assumption II.2. *The statistical relationship between a trait and any of its descendants in* $\mathbf{T}$ *is independent of the topology of* $\mathbf{T}$*.*

Assumption II.2 means that our statistical model of the evolutionary process is identical along paths through $\mathbf{T}$ from the root to any tip, and we call this the *marginal process*. It should be noted that Assumption II.2 is made for ease of exposition in this paper and that it may be relaxed, for example to model unequal rates of evolution along different branches of $\mathbf{T}$. We have assumed that $\mathbf{T}$ is a chronogram, and we will denote the *date* of a point $\mathbf{t} \in \mathbf{T}$ by the plain typeface symbol t. It is easy to see that the marginal process is Gaussian, and so it is sufficient to specify its covariance function Σ on $S \times T$, where T is the set of all dates in $\mathbf{T}$.

We call the covariance function $\Sigma_{\mathbf{T}}$, resulting from Assumptions II.1-II.2, the *phylogenetic covariance function*. It is shown in the supplement that a large class of phylogenetic covariance

functions may be constructed, offering a wide choice of priors for Bayesian evolutionary inference with function-valued traits. Some of our mathematical results are summarised in the following Proposition.

Proposition II.1. *1. If the marginal covariance function Σ is space-time separable so that it can be expressed as*

$$\Sigma((x_1,t_1),(x_2,t_2)) = K(x_1,x_2)k(t_1,t_2), \tag{5}$$

then the phylogenetic covariance function $\Sigma_{\mathbf{T}}$ is also space-time separable, i.e.

$$\Sigma_{\mathbf{T}}((x_1,\mathbf{t}_1),(x_2,\mathbf{t}_2)) = K(x_1,x_2)k_{\mathbf{T}}(\mathbf{t}_1,\mathbf{t}_2) \tag{6}$$

where $k_{\mathbf{T}}(\mathbf{t}_1,\mathbf{t}_2)$ is the phylogenetic covariance function constructed from the time-dependent component k in (5)*.*

2. When the time-dependent component k of (5) *specifies a process that is Markovian in time, we have the simple expression*

$$k_{\mathbf{T}}(\mathbf{t}_1,\mathbf{t}_2) = k(t_1,t_{12})k(t_{12},t_{12})^{-1}k(t_2,t_{12}) \tag{7}$$

where $\mathbf{t}_{12}$ is the most recent common ancestor of $\mathbf{t}_1$ and $\mathbf{t}_2$ (and t_{12} is its depth in $\mathbf{T}$).

C. Example

We close this section by constructing an example of a phylogenetic covariance function $\Sigma_{\mathbf{T}}$. Suppose that we wish to perform evolutionary inference for a set of function-valued traits. The simplest possible structure for the marginal covariance function Σ is a space-time separable function. If the traits are smooth, we should choose a spatial component which generates smooth random functions: a commonly used choice is the squared exponential covariance function:

$$K(x_1,x_2) \;=\; \exp\left(-(x_1-x_2)^2/2\theta_1^2\right). \tag{8}$$

We may believe that, conditional on any given trait, its ancestor and progenitor traits are statistically independent. This corresponds to choosing a temporal component which is Markovian. The only Markovian Gaussian processes are the class of Ornstein-Uhlenbeck processes, and have the following covariance function:

$$k(t_1,t_2) \;=\; \exp\left(-|t_1-t_2|/\theta_2\right). \tag{9}$$

From Proposition II.1 we obtain

$$\Sigma_{\mathbf{T}}((x_1,\mathbf{t}_1),(x_2,\mathbf{t}_2)) \;=\; \exp\left(-(x_1-x_2)^2/2\theta_1^2 - d_{\mathbf{T}}(\mathbf{t}_1,\mathbf{t}_2)/\theta_2\right) \tag{10}$$

where $d_{\mathbf{T}}(\mathbf{t}_1,\mathbf{t}_2) = |t_1-t_{12}|+|t_{12}-t_2|$ is the total amount of evolutionary time which elapses when moving through the chronogram $\mathbf{T}$ from $\mathbf{t}_1$ to $\mathbf{t}_2$ via their most recent common ancestor $\mathbf{t}_{12}$.

As noted above, these simple examples may be combined by summation to construct other (non-separable) phylogenetic covariance functions. As an example, in applications where there is measurement noise, the following phylogenetic covariance function, $\Sigma'_{\mathbf{T}}$, contains an uncorrelated noise term whose influence is controlled by the choice of the parameter σ_0:

$$\Sigma'_{\mathbf{T}}((x_1,\mathbf{t}_1),(x_2,\mathbf{t}_2)) \;=\; (1-\sigma_0)\exp\left(-(x_1-x_2)^2/2\theta_1^2 - d_{\mathbf{T}}(\mathbf{t}_1,\mathbf{t}_2)/\theta_2\right) + \sigma_0\delta_{\mathbf{t}_1,\mathbf{t}_2}\delta_{x_1,x_2} \tag{11}$$

where δ is the Kronecker delta.

III. ILLUSTRATIVE APPLICATION

To illustrate the above method with experimental data, we considered a version of the game 'telephone' (where players are nodes in a directed cycle graph and spoken messages are passed between successive neighbours to comic effect) with two differences: the players memorise and reproduce a drawing between neighbours, and the players are arranged as nodes in a tree graph and pass the message from root to the tips. This data was created by students from The Swinton High School and has the virtue that we know the true phylogeny and have functional data at internal nodes of the tree (including the root). We model the evolution with Eq. 11 since the data has independent errors at each point of observation, is smooth in trait-space (see e.g. the functions at the tips in Fig. 1) and, by design, may be assumed to be Markovian and stationary in evolutionary time. In the supplement we discuss the choice of hyperparameters θ_1, θ_2, σ_0. For the thirty distinct points of observation (ten measurements per function-valued trait) one uses Eq. 11 with knowledge of the phylogeny $\mathbf{T}$ and hyperparameters to construct the thirty-by-thirty matrix of covariances $\Sigma'_{\mathbf{T}}(L, L)$.

For illustration we choose the model selection task of inferring the correct tree topology, given knowledge of its branch lengths (two leaves are close, at an evolutionary separation of two time units from their common ancestor, and one leaf is more distant, at a separation of four units from the root). There are three possible trees with three tips and these branch lengths, and each choice results in a different phylogenetic covariance matrix $\Sigma'_{\mathbf{T}}(L, L)$ for use in model selection. One can use Eq. 1 to perform a simple phylogenetic inference and calculate Bayes factors for the three pairs of trees. For our choice of hyperparameters, we find Bayes factors which suggest the actual tree is a decisive choice over the other two possible trees (but note that our aim here is only to illustrate our method, and not to make a scientific claim about this particular dataset). Given the correct phylogeny, one can also calculate the posterior mean and variance of an ancestral trait using Eqs. (2,3,4). This requires the construction of $\Sigma'_{\mathbf{T}}(M, L)$ and $\Sigma'_{\mathbf{T}}(M, M)$. These summaries of the posterior are shown, for a set of points M at the root of the phylogeny, by the red and blue curves in Fig. 1B.

IV. DISCUSSION

In this paper we have exploited the powerful inference architecture provided by Gaussian processes to address phylogenetic questions for function-valued data. Explicit posterior distributions are available, giving a straightforward approach to the prediction of unobserved function-valued traits, as well as a principled approach to evolutionary model selection. The approach is suitable both for complete functional observations of function-valued traits and for discretely and even irregularly sampled traits, with missing observations.

V. APPENDIX A: EXPERIMENTS

A group of student volunteers from The Swinton High School, Salford generated the illustrative data we supply. Their desks were arranged in the form of a tree - see Fig. 2. The student positioned at the root of the tree was shown a curve, asked to commit it to memory and then, immediately after it was removed, reproduce it from memory (by tracing their finger across a tablet device). That student's drawing was then shown to his/her (downstream) neighbour in the tree who similarly committed it to memory and then attempted to reproduce it. Students located at branch nodes of the tree showed their drawings to their two neighbours in turn. In this manner

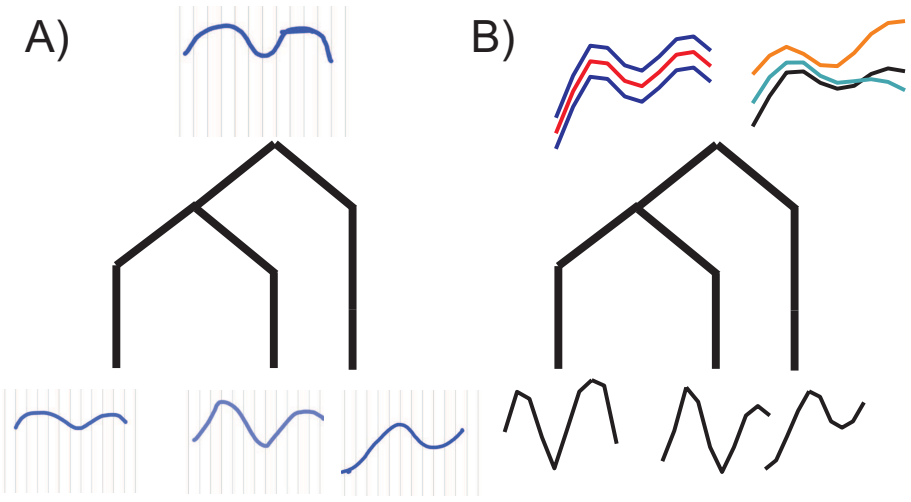

FIG. 1: Sample data generated by experiments with Class 11S1. Each piece of functional data was generated by a different student. The seating plan of the students was arranged to be in the form of a tree. We display the raw data in A) giving the curves drawn by students sitting at the root and the tips. B) Data processing and statistical inference. The black curves at the tips show the data extracted, the red curve at the root indicates the predicted mean surface given only the data from the leaves, the blue curves are one standard deviation uncertainties at each point. The three curves at the top right are example samples from the Gaussian process at the root.

the information at the root of the tree propagated to the tips.

Our aim is not to reprise existing work in hyperparameter selection (or indeed Gaussian Process Regression) but only to illustrate relevant aspects of the method we present - our approach to data analysis in this section is thus purely pedagogical. When used in earnest, maximum likelihood methods using Eq. 1 could be used to find hyperparameter choices. In the main text we consider inferring an ancestor four desks away from the tips (ignoring the curve with which the process was initialized); we suppose, by inspection of the data, that correlations in evolutionary time are appreciable. We thus choose a value for θ_2 in the Ornstein-Uhlenbek covariance (Eq. 9 of the main text) which is large: in this case thirty desks. Similarly, for the curve drawings, we observe that there are appreciable correlations on the lengthscale of around one-third of a curve. We thus set the hyperparameter θ_1 in the squared exponential covariance (Eq. 8 of the main text) to be approximately one-third of the image width. The digitisation process we used is crude and we suppose that this creates Gaussian distributed independent errors, whose variance represents $\sigma_0^2\%$ of the total variance in the data.

The curves were drawn using vertical guidelines (Fig. 2); these were removed for the purposes of analysis. Each line drawn by a student had an appreciable breadth in pixels: a mid-line was extracted. The curves were crudely aligned: they were resampled using the MATLAB 'resample' function to ensure that they all had the same number of evenly spaced points of observation, and normalized so that their minimum values were zero and maximum values were one. Our logic in this alignment was that when students copied from one image to the next they were not attempting to be accurate as regards horizontal or vertical scaling.

This experiment was part of the EPSRC funded public engagement project "Mutating Messages - A Public Experiment" EP/I017615/1 http://mutatingmessages.blogspot.com/. We would like to thank especially both the students of Class 11S1 and their teacher Ellen Pope. We would also like to thank Liam Moriarty and Karen Bultitude for their assistance in gathering this data.

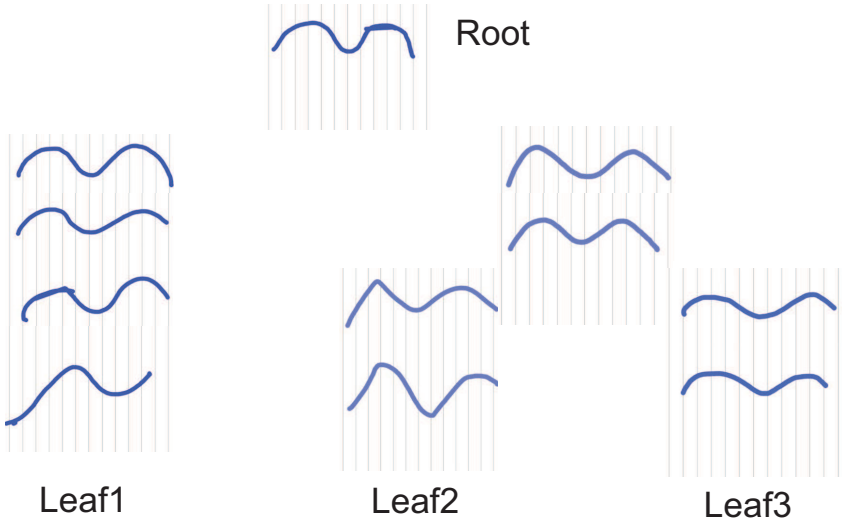

FIG. 2: Sample data generated by experiments with volunteers from Class 11S1. Each piece of functional data was generated by a different student. The seating plan of the students was arranged to be in the form of a tree; in this figure the data generated by each student is positioned to correspond to the seating plan. Essentially this was a game of 'telephone' but with participants passing on drawings, not sounds, and with the participants sitting in positions along a tree graph, not a directed cycle graph.

VI. APPENDIX B: MATHEMATICAL EXPRESSIONS FOR THE PHYLOGENETIC COVARIANCE FUNCTION

In this section we derive some mathematical expressions for the phylogenetic covariance function. Using the terminology of the main text, we first assume that the marginal covariance function is space-time separable. We do not, however, assume that the time-dependent component of the marginal covariance function is Markovian. We then give an application of these expressions to Bayesian inference for function-valued traits: given a complete functional observation, we obtain the posterior distribution of this phylogenetic Gaussian process. For necessary background on Gaussian process regression and reproducing kernel Hilbert spaces we refer to [7], sections 4.3 and 6.1.

1. Notation.

For convenience we collect here the various definitions that will be used. Again using the terminology of the main text, let

- S be the internal index set for the function-valued trait
- $\mathbf{T}$ be a phylogenetic tree whose branch lengths represent time (a chronogram), so that each

point $\theta \in \mathbf{T}$ has a time coordinate, which we will refer to as the *date* of θ. Let θ_1, θ_2 be points in $\mathbf{T}$, and let their most recent common ancestor in $\mathbf{T}$ be θ_{12}, with date t_{12}. Let t_a, t_b be respectively the earliest and latest dates of points in $\mathbf{T}$

- k be a continuous covariance function on $[t_a, t_b]$ with only finitely many nonzero eigenvalues (that is, a degenerate Mercer kernel). Let k^{12} be its restriction to $[t_a, t_{12}]$, and define

$$k_{t_1}(t_2) = k(t_1, t_2) \text{ and } k^{12}_{t_1}(t_2) = k^{12}(t_1, t_2) \tag{12}$$

Let $\mathcal{H}^{12}$ be the unique reproducing kernel Hilbert space defined by k^{12}, with inner product $\langle \cdot, \cdot \rangle_{12}$. Let ν^{12} be a Borel measure on $[t_a, t_{12}]$, and let $e^{12}_1, \ldots, e^{12}_r$ and $\lambda^{12}_1, \ldots, \lambda^{12}_r$ be respectively the eigenfunctions and (strictly positive) eigenvalues of k^{12} with respect to the measure ν^{12}, that is

$$\int_{t_a}^{t_{12}} k^{12}(s, w) e^{12}_i(w) d\nu^{12}_t(w) \;=\; \lambda^{12}_i e^{12}_i(s) \tag{13}$$

$$\int_{t_a}^{t_{12}} e^{12}_i(w) e^{12}_j(w) d\nu^{12}_t(w) \;=\; \delta(i, j) \tag{14}$$

where δ is the Kronecker delta. Then by Mercer's Theorem,

$$k^{12}(s, w) = \sum_{j=1}^{r} \lambda^{12}_j e^{12}_j(s) e^{12}_j(w) \tag{15}$$

Let Y be a Gaussian process on $[t_a, t_b]$ with covariance function k and Y^{12} be a Gaussian process on $[t_a, t_{12}]$ defined by

$$Y^{12}(w) = \sum_{i=1}^{r} \sqrt{\lambda^{12}_i} Z_i e^{12}_i(w) \tag{16}$$

where the Z_i are independent standard Normal random variables. It is easy to check that the covariance function of Y^{12} is k^{12}. The Gaussian processes Y and Y^{12} therefore have the same distribution when restricted to $[t_a, t_{12}]$. Recalling (12), for $t_{12} < t < t_b$, let $\mu^{12}_1(t), \ldots, \mu^{12}_r(t)$ be the Fourier coefficients of k_t in the basis $e^{12}_1, \ldots, e^{12}_r$:

$$\mu^{12}_i(t) = \int_{t_a}^{t_{12}} k_t(s) e^{12}_i(s) d\nu^{12}(s)$$

so that

$$k_t(s) = \sum_{i=1}^{r} \mu^{12}_i(t) e^{12}_i(s) \tag{17}$$

- $k_{\mathbf{T}}(t_1, t_2)$ be the covariance function of the phylogenetic Gaussian process on the phylogeny $\mathbf{T}$ with marginal covariance function $k(t_1, t_2)$ (this is a simple phylogenetic Gaussian process, which depends on time but has no index for the function-valued trait)

- K be a continuous covariance function on S with only finitely many nonzero eigenvalues $\lambda^S_1, \ldots, \lambda^S_n$, and corresponding eigenfunctions $e^S_1, \ldots, e^S_n$ with respect to a Borel measure ν^S.

To make this presentation self-contained, we restate here from the main text the two assumptions we use to construct a Gaussian process of function-valued traits indexed by $\mathbf{T}$:

Assumption II.1. *Conditional on their common ancestors in the phylogenetic tree* $\mathbf{T}$*, any two traits are statistically independent.*

Assumption II.2. *The statistical relationship between a trait and any of its descendants in* $\mathbf{T}$ *is independent of the topology of* $\mathbf{T}$*.*

Our main result is the following expression for the phylogenetic covariance function, which uses the Reproducing Kernel Hilbert Space inner product:

Proposition VI.1. *For* $\theta_1,\ \theta_2 \in \mathbf{T}$ *we have*

$$k_{\mathbf{T}}(\theta_1,\theta_2) \ = \ \langle k_{t_1}, k_{t_2}\rangle_{12} \tag{18}$$

$$= \ \sum_{i=1}^{r} \frac{\mu_i^{12}(t_1)\mu_i^{12}(t_2)}{\lambda_i^{12}} \tag{19}$$

Proof. Conditioning on common ancestry. *By the definition of* $k_{\mathbf{T}}$ *and the Law of Iterated Expectations we have*

$$\begin{aligned} k_{\mathbf{T}}(\theta_1,\theta_2) \ &= \ E[Y(\theta_1)Y(\theta_2)] \\ &= \ E[E[Y(\theta_1)Y(\theta_2)|Y(\theta) : \theta \text{ is an ancestor of } \theta_{12}]] \end{aligned} \tag{20}$$

Applying conditional independence. *Then by assumption 2.1,*

$$k_{\mathbf{T}}(\theta_1,\theta_2) \ = \ E[E[Y(\theta_1)|Y(\theta) : \theta \text{ is an ancestor of } \theta_{12}] \times \tag{21}$$

$$E[Y(\theta_2)|Y(\theta) : \theta \text{ is an ancestor of } \theta_{12}]] \tag{22}$$

Conditioning the marginal process. *We continue to develop* (22) *by obtaining an expression for the conditional expectation*

$$E[Y(\theta_1)|Y(\theta) : \theta \text{ is an ancestor of } \theta_{12}]$$

Lemma VI.1. *With the notation of section VI 1 we have*

$$E[Y(\theta_1)|Y(\theta) : \theta \text{ is an ancestor of } \theta_{12}] = \langle k_{t_1}, Y\rangle_{12} = \sum_{i=1}^{r} \frac{\mu_i^{12}(t)Z_i}{\sqrt{\lambda}_i^{12}} \tag{23}$$

where the final inequality holds in distribution.

Proof. Because of Assumption 2.2 and the tree structure of $\mathbf{T}$, the statement (23) is a claim about the marginal covariance function k. Then for $s \in [t_a, t_{12}]$ and $t_1 \in [t_{12}, t_b]$, since k_s^{12} is a reproducing kernel we have

$$\langle k_{t_1}, k_s^{12}\rangle_{12} = k_{t_1}(s) = k(t_1, s) \tag{24}$$

It follows from Fubini's Theorem that

$$E[\langle k_{t_1}, Y\rangle_{12} Y(s)] = E[Y(t_1)Y(s)] \tag{25}$$

and hence that the conditional expectation of $Y(t_1)$ given $\{Y(w) : w \in [t_a, t_{12}]\}$ is $\langle k_{t_1}, Y\rangle_{12}$, as required. Taking the inner product of (16) and (17) (see [7], equation (6.1)) establishes the lemma. □

To finish the proof, observe that by (21) and (23) we have

$$k_{\mathbf{T}}(\theta_1,\theta_2) \ = \ \sum_{i=1}^{r} \frac{\mu_i^{12}(t_1)\mu_i^{12}(t_2)}{\lambda_i^{12}} = \langle k_{t_1}, k_{t_2}\rangle_{12} \tag{26}$$

as required. □

A. Proof of Proposition II.1, part 1 (main text)

Let $k_{\mathbf{T}}$ be phylogenetic covariance function defined above, which depends only on time, and let X_i, $i = 1, \ldots, n$ be independent phylogenetic Gaussian processes on $\mathbf{T}$ each with covariance function $k_{\mathbf{T}}$. Define a Gaussian process by

$$f(x, \theta) = \sum_{i=1}^{n} \sqrt{\lambda_i^S} X_i(\theta) e_i^S(x) \tag{27}$$

In the terminology of the main text, it is easy to check that f is then a phylogenetic Gaussian process, indexed by both space and time, with separable covariance function

$$\begin{aligned} \Sigma_{\mathbf{T}}((x_1, \theta_1), (x_2, \theta_2)) &= k_{\mathbf{T}}(\theta_1, \theta_2) \sum_{i=1}^{n} \lambda_i e_i^S(x_1) e_i^S(x_2) && (28) \\ &= k_{\mathbf{T}}(\theta_1, \theta_2) K(x_1, x_2) && (29) \end{aligned}$$

It remains only to check that $\Sigma_{\mathbf{T}}$ has the correct marginal covariance. This follows by choosing θ_1 to be an ancestor of θ_2 in $\mathbf{T}$, since then by Assumption 2.2 we have $k_{\mathbf{T}}(\theta_1, \theta_2) = k(t_1, t_2)$ and so

$$\begin{aligned} \Sigma_{\mathbf{T}}((x_1, \theta_1), (x_2, \theta_2)) &= k(t_1, t_2) K(x_1, x_2) && (30) \\ &= \Sigma((x_1, t_1), (x_2, t_2)) && (31) \end{aligned}$$

as required. □

B. Non-separable phylogenetic covariance functions

Since the sum of any two covariance functions is again a covariance function, the space-time separable phylogenetic covariance functions obtained in section VI A allow the construction of a palette of non-separable phylogenetic covariance functions. Non-separable phylogenetic covariance functions also arise in Bayesian inference, if independent Gaussian measurement errors are added at the points of observation. A relevant discussion entitled 'making new kernels from old', including case studies, is given in [7].

C. Functional observations

Equations (2)-(4) in the main text give the posterior distribution for a phylogenetic Gaussian process after observations are made at a finite set of space-time points $L \subset S \times \mathbf{T}$. When whole function-valued traits are observed (rather than samples of function-valued traits at discrete points), we have observations at an infinite number of points in $S \times \mathbf{T}$ and so this representation no longer applies. In this case, taking the representation (27) as a Gaussian process prior enables us to form a posterior distribution. Suppose that the function-valued trait f_θ is observed at the point θ in the phylogeny. Then under the prior distribution, f_θ lies in the span of the eigenfunctions $e_1^S, \ldots, e_n^S$ and so has the representation

$$f_\theta = \sum_{i=1}^{n} a_i(\theta) e_i^S \tag{32}$$

where the $a_i(\theta)$ may be calculated by taking Fourier coefficients:

$$a_i(\theta) = \int_{w \in S} f_\theta(w) e_i^S(w) d\nu^S(w) \tag{33}$$

From the representation (27) it follows that the single functional observation $f_{\mathbf{t}}$ is equivalent to the set of point observations

$$X_i(\theta) = \frac{a_i(\theta)}{\sqrt{\lambda^S}}, \qquad i = 1, \ldots, n. \tag{34}$$

Functional observations of entire function-valued traits at $\theta_1, \ldots, \theta_p$ therefore correspond to point observations of each of the independent phylogenetic Gaussian processes $X_1, \ldots, X_n$ at each of the phylogenetic points $\mathbf{t}_j$. The posterior distributions of the Gaussian processes X_i, and hence of the original space-time Gaussian process f, then follow by Equations (2)-(4) in the main text.

D. Markovian case

We say that Σ (or, equivalently, $\Sigma_{\mathbf{T}}$) is *Markovian in evolutionary time* if, given any $\theta \in \mathbf{T}$, the function-valued traits at all ancestors of θ are statistically independent of those at all descendants of θ given the value of the trait f_θ. If Σ is time-space separable and its time-dependent component $k(t_1, t_2)$ is Markovian, formula (28) simplifies further to give the explicit expression

$$\Sigma_{\mathbf{T}}((x_1, t_1), (x_2, t_2)) = K(x_1, x_2)k(t_1, t_{12})k(t_{12}, t_{12})^{-1}k(t_2, t_{12}) \tag{35}$$

A proof may be given similar to the above (with ν^{12} the Dirac mass at t_{12}), but a simpler proof is the following. Let g be a phylogenetic Gaussian process with marginal covariance function k. Then by the Law of Iterated Expectations we have

$$k_{\mathbf{T}}(\mathbf{t}_1, \mathbf{t}_2) \;=\; E[E[g(\mathbf{t}_1)g(\mathbf{t}_2)|g(\mathbf{t}_{12})]] \tag{36}$$

By the Markov property, the functional traits at times $\mathbf{t}_1$ and $\mathbf{t}_2$ depend on their common ancestry only through the Normal random variable $g(\mathbf{t}_{12})$. Assumption II.1 therefore gives that

$$k_{\mathbf{T}}(\mathbf{t}_1, \mathbf{t}_2) \;=\; E[E[g(\mathbf{t}_1)|g(\mathbf{t}_{12})]E[g(\mathbf{t}_2)|g(\mathbf{t}_{12})]] \tag{37}$$

By Assumption II.2 and standard properties of Normal random variables we have

$$E[g(\mathbf{t}_1)|g(\mathbf{t}_{12})] = k(t_1, t_{12})k(t_{12}, t_{12})^{-1}g(\mathbf{t}_{12}) \tag{38}$$

from which the required result follows, after substitution into (37). □

[1] Meyer, K. & Kirkpatrick, M. 2005 Up hill, down dale: quantitative genetics of curvaceous traits *Phil. Trans. R. Soc. B* **360**(1459), 1443-1455. (DOI 10.1098/rstb.2005.1681.)

[2] Ramsay, J. O. & Silverman, B.W. 2005 *Functional data analysis.* Springer.

[3] Wiens, J.J. 2001 Character analysis in morphological phylogenetics: problems and solutions. *Syst Biol.* **50**(5), 689-99.

[4] Felsenstein, J. 1985 Phylogenies and the Comparative Method. *Am. Nat.* **125**(1), 1-15.

[5] Martins, E.P. & Hansen, T.F. 1997 Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *Am. Nat.* **149**(4), 646-667.

[6] Kingsolver, J.G., Gomulkiewicz, R. & Carter, P.A. 2001 Variation, selection and evolution of function-valued traits. *Genetica* **112-113**, 87-104. (DOI 10.1023/A:1013323318612.)

[7] Rasmussen, C.E. & Williams, C.K.I. 2006 *Gaussian processes for machine learning.* MIT Press.

[8] Stein, M.L. 1999 *Interpolation of spatial data: some theory for kriging.* Springer.

[9] Patil, G.P. & Rao, C.R. 1993 *Multivariate environmental statistics.* North-Holland.

[10] Grafen, A. 1989 The Phylogenetic Regression. *Proc. R. Soc. B* **326**(1233), 119-157.

[11] Macholan, M. 2008 The mouse skull as a source of morphometric data for phylogeny inference. *Zool. Anz.* **247**(4), 315-327.